The Double-Edged Sword of AI: Evidence on Student Engagement from the Classroom*

Jaime Polanco-Jiménez† Kristof De Witte‡

August 19, 2025

Abstract

We evaluate how generative-AI chatbot design affects student engagement and learning in a randomized trial with 2,440 secondary students. We find chatbot design has a powerful, non-monotonic effect on student engagement: assignment to a curriculum-tailored chatbot increased module completion by 9.7 percentage points, while a generic chatbot decreased it by 5.0 points. This engagement effect is the primary mechanism for learning. Our most robust causal estimate is the intent-to-treat (ITT) effect of the offer of the tailored chatbot, which increased immediate test scores by 0.06 standard deviations on the full sample under conservative imputation assumptions. For the "compliers" induced to complete the module by the tailored design, the effect on learning is larger and more durable, increasing knowledge retention two months later by 0.36 standard deviations. Curricular integration appears critical for solving the first-order problem of engagement in educational technology.

JEL Codes: I21, C93, G51, O33

Keywords: Student Engagement, Educational Technology, Chatbots, Curricular Integration, Randomized Controlled Trial, Knowledge Retention, Human Capital

*Authors acknowledge financial support from the Horizon Europe project BRIDGE (grant 211012702). Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

†Corresponding author. Leuven Economics of Education Research, KU Leuven, Belgium; Department of Economics at Pontificia Universidad Javeriana, Colombia (email: jaime.polancojimenez@kuleuven.be, jaime.polanco@javeriana.edu.co)

‡Full Leuven Economics of Education Research, KU Leuven, Belgium; UNU-Merit, Maastricht University, the Netherlands (email: kristof.dewitte@kuleuven.be)

1 Introduction

School systems worldwide are grappling with persistent teacher shortages (Pressley, 2021; Sutcher, Darling-Hammond, & Carver-Thomas, 2019) and are increasingly turning to Artificial Intelligence (AI) as a scalable solution. Yet they face a first-order question with no causal evidence: Is it better to adopt widely available, general-purpose generative AI chatbots, or must tools be deeply tailored to the local curriculum? The answer is critical. While the literature establishes that instructional time is a key input to the educational production function (Aucejo & Romano, 2016; Jaume & Willén, 2019), another line of research shows that the effectiveness of educational technology hinges on its thoughtful implementation and integration with the curriculum (Bai, Mo, Zhang, Boswell, & Rozelle, 2016; Cacault, Hildebrand, Laurent-Lucchetti, & Pellizzari, 2021; Figlio, Rush, & Yin, 2013). This paper provides the first large-scale experimental evidence to resolve this tension in the context of modern AI.

We find that curricular integration is the key ingredient for success, operating through a powerful effect on student engagement. Our randomized trial shows that chatbot design is a double-edged sword: assignment to a curriculum-tailored chatbot increased the probability of students completing a learning module by 9.7 percentage points, whereas assignment to a generic chatbot *decreased* completion by 5.0 percentage points relative to traditional instruction. This non-monotonic effect on engagement is our central finding and the primary mechanism driving our learning outcomes. While recent meta-analyses suggest AI tools generally have a positive impact (Tlili, Saqer, Salha, & Huang, 2025; Wang et al., 2024; Wu & Yu, 2023), our results provide a crucial qualification: poor design can be actively harmful to student participation, a finding that contrasts with the optimistic view in much of the current literature (e.g., Henkel, Horne-Robinson, Kozhakhmetova, & Lee, 2024; Kestin, Miller, Klales, Milbourne, & Ponti, 2024).

To generate this evidence, we conduct an RCT with 2,440 Belgian secondary students to evaluate the impact of chatbot design on financial literacy—a critical form of human capital

where knowledge gaps are wide and consequential (Lusardi & Mitchell, 2014). We randomly assign students within classrooms to one of three arms: traditional instruction (Control), a 'Generic Chatbot' using a general knowledge base, and a 'Tailored Chatbot'. Our tailored chatbot combines two key features: content-specificity (trained on the Belgian tax code) and pedagogical adaptivity (offering personalized feedback). Our design therefore estimates the joint effect of these two 'tailoring' dimensions, providing a crucial test of the returns to contextualization.

The differential engagement we document translates directly into learning. Our most robust estimate is a modest but significant Intent-to-Treat (ITT) effect on the full sample: offering the tailored chatbot increased immediate test scores by 0.036 standard deviations. However, the learning effects are far more pronounced for the specific students whose behavior was changed by the intervention. For the "compliers" induced to complete the module by the tailored chatbot's superior design, the treatment increased knowledge retention two months later by a significant 0.36 standard deviations, addressing the "summer slide" or fade-out problem common to many educational interventions (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996). A key channel for this success appears to be student self-perception: our results suggest the tailored chatbot boosts student self-confidence, while the generic tool harms it.

This study makes three contributions. First, by directly comparing a generic and a tailored chatbot, we provide the first causal estimates on the returns to contextualization for generative AI in education, updating the seminal findings of Bai et al. (2016) for the modern AI era. Second, we broaden the set of outcomes for evaluating educational technology, documenting powerful effects on student engagement and, crucially, on long-term knowledge retention, a dimension often overlooked in the literature. Third, we provide evidence on the mechanisms driving these effects. The tailored chatbot's success is mediated not just through direct instruction, but through its ability to solve the first-order problem of student engagement and to bolster self-confidence, offering a glimpse inside the "black box" of AI-

driven learning.

The remainder of this paper proceeds as follows. Section 2 details the compelling institutional context of financial education in Flanders, highlighting a paradox of high average student performance coupled with deep socioeconomic inequality that motivates our intervention. Section 3 describes our experimental design, the rich baseline and longitudinal data collected, and presents a detailed diagnosis of the differential attrition that is a central finding of our paper. Building on this, Section 4 lays out our empirical strategy, which prioritizes robust Intent-to-Treat (ITT) estimates, supplemented by non-parametric bounds, and a carefully specified Instrumental Variable (IV) model to estimate the Local Average Treatment Effect (LATE) for a specific subgroup. Section 5 presents our main results in a structured sequence, beginning with the powerful, non-monotonic effect of chatbot design on student engagement, then presenting the robust ITT and LATE estimates for learning outcomes, and finally exploring potential mechanisms. Section 6 concludes by discussing the implications of our findings for the design of effective educational technology and future policy.

2 Financial Education in Flanders: A Case for Tailored AI

Financial literacy is a critical form of human capital with substantial, long-run consequences for household economic security (Lusardi & Mitchell, 2014). In response, a growing number of countries have integrated financial education into their secondary school curricula (OECD, 2020). Belgium's Flemish community offers a compelling setting to study the implementation of this mandate. Following a major 2019 reform, financial literacy became a key cross-curricular competence for all secondary students, as mandated by the region's educational modernization act (Vlaamse Regering, n.d.).

This policy, however, has produced a paradox: while Flemish students rank among

the world's best on PISA financial literacy assessments, this high average masks a severe achievement gap linked to socioeconomic status (De Witte, De Beckker, & Holz, 2020). The most recent PISA data confirm this stark inequality. In 2022, the performance gap between socio-economically advantaged and disadvantaged students in Flanders was 104 score points, substantially larger than the OECD average of 87 points (OCDE, 2024). Even more telling, a student's economic, social, and cultural status (ESCS) explains 16.8% of the variance in financial literacy performance in Flanders—one of the strongest such relationships among developed economies and far exceeding the OECD average of 11.6%.[1] This evidence underscores that while average performance is high, the educational system in Flanders struggles to decouple academic achievement from students' family backgrounds, creating a clear policy imperative for interventions that can deliver high-quality, standardized instruction to all students.

This inequality stems from two core implementation challenges documented by De Witte et al. (2020). First, as a cross-curricular subject, financial literacy is often taught by non-specialists who may lack deep content knowledge, creating a demand for standardized, high-quality instructional resources. Second, Flemish classrooms exhibit significant student heterogeneity across academic (ASO), technical (TSO), and vocational (BSO) tracks, making a one-size-fits-all approach ineffective and "differentiated instruction" a policy priority. These dual needs—for standardization to ensure quality and for personalization to address heterogeneity—present a fundamental tension for policymakers. Our experiment is designed to test whether a single intervention, a curriculum-tailored AI tool, can resolve this tension by providing both standardized content and adaptive, personalized instruction.

Our study is situated at the intersection of literatures on the educational production function, educational technology, and AI. A long line of research shows that simply increasing instructional time often yields surprisingly small returns (Hanushek, 2003; Jaume & Willén, 2019). Instead, the quality and efficiency of that time are the primary drivers of learning

[1]The PISA index of economic, social and cultural status (ESCS) is a composite measure derived from student reports on parental occupations, parental education, and home possessions.

(Aucejo & Romano, 2016). This insight motivates our focus on outcomes beyond immediate test scores. We analyze learning efficiency—the knowledge gained per unit of time—and the durability of learning, or long-term knowledge retention. Demonstrating a lasting impact is particularly important, as many educational interventions tend to lose their effectiveness over time (Cooper et al., 1996).

Technology is often proposed as a solution to enhance instructional quality at scale, but evidence on its effectiveness is mixed. While some studies find positive effects, rigorous experimental evaluations often find null or even negative impacts from simply replacing in-person teaching with standard online formats (Cacault et al., 2021; Figlio et al., 2013). This suggests that implementation details are paramount. Indeed, the seminal work of Bai et al. (2016) shows causally that the effectiveness of ICT depends critically on its integration with the local curriculum.

The latest wave of EdTech, powered by AI, promises to overcome the limitations of older online tools by offering personalized and adaptive learning. Recent meta-analyses confirm that AI-powered tools can have a positive effect on student learning (e.g., Tlili et al., 2025; Wang et al., 2024; Wu & Yu, 2023). However, the rise of powerful, general-purpose AI models introduces a new dimension to the finding of Bai et al. (2016). The question is no longer simply whether to integrate technology, but how deeply. Is it sufficient to use a generic AI tool that understands a topic broadly (our T1 arm), or is the key to unlocking educational productivity to use an AI that is deeply tailored to the specific local curriculum (our T2 arm)? To our knowledge, no large-scale randomized trial has causally estimated the differential returns to generic versus curriculum-tailored AI. This study is designed to fill this critical gap.

This institutional context and literature motivate a clear set of testable hypotheses. First, consistent with recent meta-analyses (e.g., Wu & Yu, 2023), we expect both AI interventions to improve learning outcomes relative to traditional instruction. Second, and central to our contribution, we test the returns to contextualization. Motivated by the criti-

cal role of curricular integration (Bai et al., 2016), we hypothesize that the tailored AI (T2) will be significantly more effective than its generic counterpart (T1). Third, we predict the primary advantages of the tailored AI will be in improved learning efficiency and superior long-term knowledge retention, addressing the fade-out problem common to many interventions (Cooper et al., 1996), rather than in immediate test score gains. Finally, we explore mechanisms, positing that the tailored AI's success is mediated by its positive impact on non-cognitive outcomes, specifically by fostering greater student engagement (proxied by higher completion rates) and enhancing academic self-confidence (cf. Sales & Pane, 2020).

We also acknowledge two potential threats to the generalizability of our findings. First, our choice of topic, taxes, is one where students may have strong pre-existing beliefs. Second, the effectiveness of the chatbots could depend on students' prior attitudes toward technology. Our rich baseline data allow us to test these hypotheses directly. In Section 5.2, we present a formal heterogeneity analysis and show that our main engagement effects are remarkably stable across these dimensions of student attitudes, strengthening the external validity of our conclusions.

3 Data and Experimental Context

We assess the impact of generative AI chatbot design on student engagement and learning through a large-scale randomized controlled trial (RCT) conducted from January to May 2025 in the Flemish secondary school system in Belgium.[2] Our study population consists of 2,440 students in their third grade of secondary school (typically ages 16-18) from 120 classrooms across 58 schools. The experiment was embedded within the standard curriculum on Economic and Financial Literacy, focusing on the complex Belgian personal income tax system designed for Educational Master in Economics in KU Leuven.

Upon enrollment, students were randomly assigned at the individual level, within class-

[2]This trial was pre-registered in the AEA RCT Registry on January 27, 2025, with ID AEARCTR-0015266. The pre-analysis plan is available at `https://doi.org/10.1257/rct.15266-1.0`.

rooms, to one of three experimental arms. This design non-parametrically controls for unobserved heterogeneity across teachers, classrooms, and peer groups. Students in the control group (T0) followed the traditional learning path, using existing course materials. The first treatment group, the 'Generic Chatbot' (T1) arm, received condensed instruction supplemented by a chatbot with general knowledge of taxation principles but no specific details of the Belgian tax code. Finally, students in the 'Tailored Chatbot' (T2) arm engaged with an adaptive chatbot designed specifically for the Flemish curriculum and the Belgian tax code. This tool combines two key features: content-specificity and pedagogical adaptivity, personalizing the learning path based on student responses. Our design therefore estimates the combined effect of these two 'tailoring' dimensions.

3.1 Data Collection and Variable Construction

We collected data via online questionnaires at three points in time: a pre-test (t=0), an immediate post-test (t=1), and a follow-up test two months later (t=2). Our analysis examines three categories of outcomes: learning outcomes, psychosocial outcomes, and a descriptive measure of efficiency.

Our primary learning outcomes are twofold. First, **Gained Financial Literacy** measures immediate learning, calculated as the difference between a student's post-test and pre-test score. Second, **Knowledge Retention** measures the persistence of learning, using the student's score on the follow-up test. To ensure comparability, questions for all three test waves were developed from a common item bank of 10 multiple-choice items and validated for equivalent difficulty by subject-matter experts. Both learning outcomes are standardized using the control group's distribution for ease of interpretation.

Second, we analyze the treatment's impact on a range of **Psychosocial Outcomes**. We administered a comprehensive battery of psychosocial constructs at both pre-test and post-test, allowing us to measure the change in these dimensions as an outcome. These instruments were adapted from seminal, validated scales in the educational psychology liter-

ature, including measures of *Attitude & Motivation* from the MSLQ (Pintrich, Smith, Garcia, & McKeachie, 1991), *Self-Confidence* from the General Self-Efficacy Scale (Schwarzer & Jerusalem, 1995), and *Engagement* from an adaptation of the Utrecht Work Engagement Scale (Schaufeli, Salanova, González-Romá, & Bakker, 2002). A full list of the constructs and their sources is detailed in Appendix A.

The pre-treatment collection of this rich data also serves three critical functions for our identification strategy. The baseline measures of demographics, prior academic achievement, and these same psychosocial constructs are used to: (1) conduct a comprehensive balance check to validate our randomization; (2) enable a detailed diagnosis of the selection into attrition, which is our key mechanism; and (3) facilitate a robust exploration of heterogeneous treatment effects across key student subgroups.

Finally, we constructed a descriptive measure of **Learning Efficiency**. This is defined for the subsample of module completers as their standardized knowledge gain divided by the time spent on the module, as shown in Equation 1. Time was logged by the online platform as the total duration between starting and submitting the module, a potentially noisy proxy for active learning time. Because this variable is undefined for the 70% of students who attrited, it cannot be used in our primary causal analyses. We therefore use it only for descriptive purposes.

$$\text{Learning Efficiency}_i = \frac{\text{Standardized Gained Financial Literacy}_i}{\text{Time Spent on Module (minutes)}_i} \tag{1}$$

3.2 Baseline Balance of the Randomized Sample

We first verify that our randomization produced statistically equivalent groups across the full sample prior to the intervention. Table 1 presents the means and standard deviations of baseline characteristics for each experimental arm, along with p-values for the difference between each treatment group and the control group. The balance for categorical variables is shown in Appendix Table 7. The tables show no statistically significant differences at

conventional levels across dozens of pre-determined characteristics. This comprehensive evidence confirms that the randomization was successful, providing a strong foundation for our causal analysis.

Table 1: Baseline Balance Check: Continuous Variables

Variable	(1) Control Mean (SD)	(2) Generic AI Mean (SD)	(3) Tailored AI Mean (SD)	(4) p-val (T1-Ctrl)	(5) p-val (T2-Ctrl)
Pre-Intervention Outcomes					
Financial Literacy Score (Pre-Test)	0.349	0.337	0.343	0.451	0.723
	(0.242)	(0.246)	(0.238)		
Psychosocial Scales (1-5)					
Attitude and Motivation	2.866	2.829	2.796	0.315	0.108
	(0.733)	(0.735)	(0.749)		
Learning & User Experience	2.781	2.781	2.749	0.998	0.452
	(0.888)	(0.879)	(0.845)		
Self-Regulation & Metacognition	2.712	2.647	2.669	0.104	0.298
	(0.821)	(0.780)	(0.781)		
Engagement & Commitment	2.548	2.491	2.507	0.127	0.301
	(0.776)	(0.711)	(0.779)		
Self-Confidence & Self-Efficacy	2.683	2.686	2.693	0.932	0.814
	(0.867)	(0.837)	(0.866)		
Emotional & Psychological Factors	2.884	2.857	2.897	0.421	0.763
	(0.684)	(0.667)	(0.706)		
Observations	799	870	771		

Notes: This table reports means of continuous baseline characteristics for the full randomized sample (N=2,394). Standard deviations are in parentheses. Columns 4 and 5 report p-values from OLS regressions of each baseline characteristic on treatment indicators for the Generic AI (T1) and Tailored AI (T2) groups, respectively, with the Control group as the omitted category. Regressions include school fixed effects. Standard errors are robust and clustered at the school level (58 clusters). No p-value is significant at the 10% level, providing strong evidence of successful randomization.

3.3 The Central Empirical Challenge: High and Differential Attrition

While the full sample was balanced at baseline, a central feature of our experiment is a high overall rate of attrition that varies starkly across treatment arms, making it a key economic outcome. Figure 1 documents the participant flow. Of the 2,440 randomized students, only 617 (25.3%) provided complete post-test data.

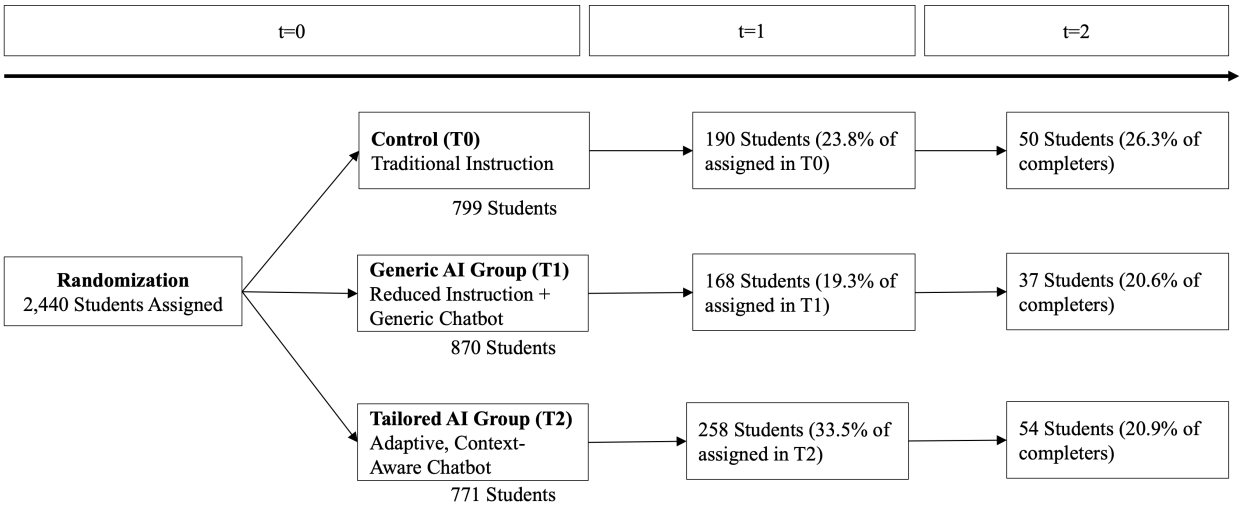

Figure 1: Experimental Design and Participant Flow

Notes: This figure shows the flow of participants through the randomized controlled trial. Numbers in boxes represent the count of participants at each stage. The sample at t=0 represents the full randomized sample. Completion rates at the post-test (t=1) are calculated relative to the number of students initially assigned to that arm. The follow-up completion rate at t=2 is calculated relative to the number who completed the post-test at t=1.

This attrition is not random across groups. The completion rate for the tailored chatbot group (T2) was 33.5%, substantially higher than the control group's 19.3%. In contrast, the generic chatbot group (T1) had the lowest completion rate at just 23.8%.

This differential attrition is a first-order finding of our paper: the tailored intervention was significantly more engaging than traditional instruction, while the generic tool actively disengaged students. This implies that simple comparisons on the sample of completers would yield biased estimates and motivates an empirical strategy, detailed in Section 4, that relies on Intent-to-Treat (ITT) estimates to recover credible causal parameters.

3.4 Diagnosing Selection: Who Attrites?

Given the differential attrition, we now diagnose the selection mechanism by examining which students persist to completion within each arm. Table 2 compares the baseline characteristics of students who completed the module versus those who did not.

Table 2: Diagnosing Selection: Baseline Characteristics of Completers vs. Non-Completers

Variable	Control (T0)		Generic AI (T1)		Tailored AI (T2)	
	Completer	Non-Comp.	Completer	Non-Comp.	Completer	Non-Comp.
Pre-Test Score	0.371 (0.240)	0.318 (0.253)	0.385 (0.236)	0.306 (0.254)	0.354 (0.230)	0.322 (0.252)
Attitude & Motivation	2.903 (0.674)	2.822 (0.740)	2.918 (0.797)	2.809 (0.719)	2.800 (0.693)	2.765 (0.748)
AI Attitude & Motivation	2.988 (1.000)	2.824 (0.925)	2.988 (0.943)	2.951 (0.962)	3.016 (0.938)	2.861 (0.916)
Learning Experience	2.974 (0.919)	2.704 (0.862)	2.895 (0.895)	2.756 (0.887)	2.750 (0.839)	2.726 (0.824)
Self-Regulation	2.723 (0.801)	2.698 (0.823)	2.646 (0.759)	2.634 (0.784)	2.672 (0.726)	2.645 (0.812)
Engagement & Commitment	2.584 (0.771)	2.528 (0.777)	2.519 (0.761)	2.478 (0.696)	2.513 (0.735)	2.476 (0.794)
Self-Confidence	2.738 (0.811)	2.663 (0.890)	2.855 (0.971)	2.633 (0.784)	2.689 (0.795)	2.696 (0.896)
Emotional Factors	2.926 (0.655)	2.886 (0.687)	2.900 (0.663)	2.838 (0.674)	2.870 (0.707)	2.887 (0.713)
Observations	235	564	201	669	300	471

Note: Table reports means with standard deviations in parentheses. It diagnoses the selection into module completion by comparing the baseline characteristics of students who completed the post-test versus those who did not, within each treatment arm. Statistical significance of the differences is discussed in the text.

The analysis reveals distinct selection patterns across the three arms, with formal statistical tests provided in Appendix Table 8. In the Control group, we observe broad positive selection: students who persisted began with significantly higher pre-test scores and more positive attitudes.

Assignment to the Generic Chatbot exacerbates this selection on academic ability and reshapes the role of psychosocial factors. The gap in pre-test scores between completers and non-completers is at its largest and most statistically significant in this arm ($p < 0.001$). Furthermore, the single strongest psychosocial predictor of completion is baseline self-confidence, where completers score significantly higher ($p = 0.007$). This paints a clear picture: the generic tool appears to be so unhelpful that only the academically strongest and most self-confident students persist, while others are driven to attrite.

In contrast, the Tailored Chatbot fundamentally alters this selection process. Most importantly, it mitigates the strong selection on prior academic ability; the difference in pre-test scores between completers and non-completers is substantially smaller and only marginally significant ($p = 0.082$). The tailored tool appears to democratize engagement, making it accessible beyond just the highest-achieving students. Instead of academic ability or innate self-confidence, the primary psychosocial predictor of completion in this arm becomes a student's baseline attitude toward AI ($p = 0.032$). This suggests that when a tool is well-designed and effective, the main determinant of its use is simply a student's willingness to

engage with the medium itself.

4 Empirical Strategy

Our empirical strategy is designed to identify the causal effects of different Generative-AI chatbot designs on student outcomes in the presence of high and differential attrition. We proceed in four stages. First, we define the ideal experiment and the estimands of interest within the framework of potential outcomes. Second, we articulate how non-random attrition challenges the identification of these estimands. Third, we detail our primary estimation strategy, which focuses on the Intent-to-Treat (ITT) effect, supported by non-parametric bounds to ensure robustness. Finally, we specify an Instrumental Variable (IV) model to estimate the Local Average Treatment Effect (LATE) for the subpopulation of students whose engagement was positively affected by the tailored AI intervention.

4.1 Conceptual Framework and the Ideal Experiment

Following the potential outcomes framework (Angrist & Pischke, 2009a, Ch. 2), we define a set of potential outcomes for each student i. Let $Y_i(k)$ be the potential outcome (e.g., test score) for student i if assigned to experimental arm k, where $k \in \{0, 1, 2\}$ for the Control, Generic AI (T1), and Tailored AI (T2) groups, respectively. The causal effect of a given treatment relative to the control is the difference in potential outcomes, e.g., $Y_i(2) - Y_i(0)$. The primary parameter of interest is the Average Treatment Effect (ATE), defined as $ATE_k = E[Y_i(k) - Y_i(0)]$.

In an ideal experiment with full compliance and no attrition, the random assignment ($Z_{is,k}$) of student i in school s to treatment k would be equivalent to actual treatment completion ($T_{is,k}$). Due to randomization, the assignment would be independent of potential outcomes, i.e., $\{Y_i(0), Y_i(1), Y_i(2)\} \perp Z_{is,k}$. A simple comparison of mean outcomes across arms would then consistently estimate the ATE.

4.2 The Identification Challenge: High and Differential Attrition

Our experiment deviates from this ideal due to substantial and differential attrition, as documented in Section 3.3. This non-compliance breaks the simple link between assignment and treatment. A naive comparison of outcomes for only those students who completed the module would be subject to severe selection bias (Angrist & Pischke, 2009a, Ch. 3.2). The students who persist in each arm are not random subsets of their original assignment groups; they are self-selected based on observed and unobserved characteristics that are likely correlated with potential outcomes. For example, if only the most motivated students persist in the Generic AI arm, a comparison of completers would mistakenly attribute their high performance to the AI tool rather than their pre-existing motivation. This selection problem necessitates an empirical strategy that relies on the initial random assignment for identification.

As we demonstrate in a detailed diagnostic analysis in Appendix D, this selection bias is not merely a theoretical concern; it is empirically massive. For instance, the naive estimate of the effect of the tailored chatbot on long-term retention is nearly 0.4 standard deviations, while our more credible causal estimates are close to zero. This large discrepancy provides a stark illustration of the selection problem and necessitates an empirical strategy that relies on the initial random assignment for identification.

4.3 Primary Estimand: Intent-to-Treat (ITT) Effects and Lee Bounds

Our primary analysis focuses on estimating the Intent-to-Treat (ITT) effect on our key learning outcomes. The ITT is the causal effect of being *offered* the treatment, regardless of module completion, and is a highly policy-relevant parameter as it measures the overall impact of a program rollout. We estimate the ITT on the full randomized sample (N=2,440) using the following specification:

$$Y_{is} = \alpha + \delta_1 Z_{is,1} + \delta_2 Z_{is,2} + \mathbf{X}'_{\mathbf{is}}\gamma + \mu_s + \eta_{is} \tag{2}$$

where Y_{is} is the outcome for student i in school s for instance, it represents the gained financial literacy for student i in school s, calculated as their post-test score minus their pre-test score. $Z_{is,k}$ is a dummy variable equal to 1 if the student was assigned to treatment arm k, $\mathbf{X_{is}}$ is a vector of baseline student characteristics, and μ_s are school fixed effects.

A key challenge in estimating Equation 2 is that the post-test score, and therefore the gain score Y_{is}, is missing for the 74% of students who attrited. Our main specification addresses this by imputing a knowledge gain of zero for all attritors.

The economic rationale for this choice is the assumption of a *weakly monotonic treatment effect on knowledge*. That is, exposure to a curriculum-aligned learning module, even if incomplete or frustrating, is unlikely to cause a net loss of knowledge relative to a student's baseline. The modules are designed to augment a student's stock of human capital; the most plausible worst-case scenario for a student who disengages is that they learned nothing from the experience, resulting in zero knowledge gain. By imputing a value of zero, we adopt this most pessimistic scenario for every attritor. This ensures that our ITT estimate is a highly conservative lower-bound on the true average treatment effect, a standard approach in experimental analysis (Angrist & Pischke, 2009b; Duflo, Glennerster, & Kremer, 2007; Kling, Liebman, & Katz, 2007). To confirm that our findings are not dependent on this specific choice, we also estimate non-parametric bounds on the Average Treatment Effect (ATE) following Tauchmann (2014), which provides a sensitivity case bounds under a similar monotonicity assumption about selection.

4.4 The Effect of Completion: A LATE Framework for the Tailored AI

While the ITT provides a policy-relevant population average, we are also interested in the effect of actually completing the AI-supported learning module. We use an Instrumental Variable (IV) strategy to estimate this parameter, identifying the Local Average Treatment Effect (LATE), which is the average treatment effect for the specific subpopulation of students whose completion behavior is changed by the random assignment (Angrist & Pischke, 2009a, Ch. 4.4).

The interpretation of the LATE parameter is critically dependent on the nature of the first-stage relationship—that is, whether the offer of a treatment encourages or discourages completion (see also Cunningham, 2018, Ch. 9). If the instrument encourages participation (a positive first stage), the LATE identifies the causal effect for "compliers": students who complete the module only because they were offered that specific intervention. If the instrument discourages participation (a negative first stage), the standard complier group does not exist in a meaningful way, and the LATE parameter is not interpretable as the effect for a policy-relevant group. In our Results section, we will estimate the first-stage effects for both chatbots and apply the appropriate interpretation.

We estimate this parameter using a Two-Stage Least Squares (2SLS) model. While our model includes two endogenous variables (completion of T1 and T2), here our parameters of interest are β_1 and β_2:

$$\textit{First Stages:} \quad T_{is,k} = \pi_{k0} + \pi_{k1} Z_{is,1} + \pi_{k2} Z_{is,2} + \mathbf{X'_{is}}\omega_k + \mu_s + \nu_{is,k} \quad \text{for } k = 1, 2 \tag{3}$$

$$\textit{Second Stage:} \quad Y_{is} = \beta_0 + \beta_1 \hat{T}_{is,1} + \beta_2 \hat{T}_{is,2} + \mathbf{X'_{is}}\lambda + \mu_s + \eta_{is} \tag{4}$$

Here, $T_{is,k}$ is an indicator for student i completing the module in arm k, instrumented by the assignment indicators $Z_{is,k}$. The validity of this IV strategy rests on three key assumptions.

First, Instrument Relevance, requires that random assignment strongly predicts treat-

ment completion. As we show in Table **??**, our first-stage F-statistics are exceptionally large.

Second, Monotonicity, requires that the instrument pushes all students' participation decisions in the same direction (i.e., there are no "defiers" who would complete the module only if assigned to the control, but not if assigned to the treatment).

Third, the *Exclusion Restriction*, requires that random assignment affects a student's outcome only through its effect on their completion of the learning path. The primary threat would be a direct psychological effect of the assignment itself. We argue this is unlikely to be a first-order concern, as the intensive learning module is a far more substantial treatment than the simple knowledge of one's assignment. Moreover, our within-classroom randomization non-parametrically controls for any general Hawthorne effects common to all students in the experiment, strengthening the credibility of this assumption.

5 Results

Our analysis proceeds in a structured sequence to build a cohesive narrative. We begin by presenting our main finding: the causal effect of generative-AI chatbot design on student engagement. We then explore the heterogeneity of this engagement effect. Next, we present our primary causal estimates for learning outcomes using an Intent-to-Treat (ITT) framework with a comprehensive sensitivity analysis. We follow this with a secondary analysis estimating the Local Average Treatment Effect (LATE). Finally, we explore potential mechanisms by examining psychosocial outcomes on the selected sample of completers.

5.1 Main Finding: Chatbot Design Powerfully Affects Student Engagement

The first and most direct impact of our intervention is on whether students persist through the learning module. Table 3 presents the Intent-to-Treat (ITT) effect of treatment assign-

ment on the probability of completing the post-test. The results show a stark divergence driven by chatbot design. Assignment to the tailored AI (T2) significantly increased the completion rate by 9.7 percentage points ($p < 0.01$) relative to the control group. In dramatic contrast, assignment to the generic AI (T1) *decreased* the completion rate by 5.0 percentage points ($p < 0.01$). This non-monotonic effect is our paper's central finding: curricular integration is the key determinant of student engagement, while a non-contextualized tool can actively disengage students.

Table 3: The Effect of Treatment Assignment on Module Completion (First Stage)

Dependent Variable:	(1) Completed Post-Test (0/1)	(2) Completed Post-Test (0/1)
Assigned to Generic AI (T1)	-0.0504***	-0.0447**
	(0.0058)	(0.0211)
Assigned to Tailored AI (T2)	0.0973***	0.0968***
	(0.0119)	(0.0217)
Control Group Mean	0.238	0.426
Observations	2,440	2,440
R-squared	0.045	0.018
Model Specification		
Baseline Controls	Yes	No
School Fixed Effects	Yes	No

Notes: This table reports estimates of the Intent-to-Treat (ITT) effect on the probability of completing the post-test. The sample is the full set of randomized students (N=2,440). The coefficients represent the effect of treatment assignment relative to the control group (the omitted category). Column (1) includes school fixed effects and a full set of baseline controls (pre-test score, gender, parental education, prior grades). Column (2) is a parsimonious specification without controls. Robust standard errors, clustered by school, are in parentheses. The Control Group Means represents the raw completion rate for the control group.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.2 Heterogeneous Engagement Effects by Student Attitudes

Having established the average effect of chatbot design on engagement, we now examine whether this effect is consistent across the student population or is driven by specific subgroups. Table 4 explores this by interacting treatment assignment with three key baseline

characteristics: students' perception of taxes, their general attitude toward AI, and their dominant learning style based on the Honey-Mumford framework (Honey & Mumford, 1986).

Table 4: Heterogeneous Effects of Chatbot Assignment on Module Completion

Dependent Variable:	Completed Post-Test (0/1)		
Model:	(1) By Tax Perception	(2) By AI Attitude	(3) By Learning Style
Treatment Main Effects			
Assigned to Generic AI (T1)	-0.032	-0.051**	0.013
	(0.027)	(0.021)	(0.040)
Assigned to Tailored AI (T2)	0.094***	0.096***	0.152***
	(0.034)	(0.028)	(0.052)
Interaction Effects			
T1 × Negative Tax Perception	-0.039		
	(0.040)		
T2 × Negative Tax Perception	0.007		
	(0.046)		
T1 × Standardized AI Attitude		-0.030	
		(0.019)	
T2 × Standardized AI Attitude		0.006	
		(0.023)	
T1 × Activist			-0.033
			(0.050)
T2 × Activist			-0.035
			(0.062)
T1 × Pragmatist			-0.111*
			(0.058)
T2 × Pragmatist			-0.090
			(0.068)
T1 × Reflector			-0.033
			(0.052)
T2 × Reflector			0.051
			(0.070)
Controls & Fixed Effects			
Baseline Controls	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes
Observations	2,350	2,301	2,413
R^2	0.045	0.046	0.457

Notes: This table reports heterogeneous treatment effects on the probability of completing the post-test. Each column interacts treatment assignment with a different baseline characteristic. Column (1) interacts with a binary indicator for negative tax perception. Column (2) interacts with a standardized measure of baseline AI attitude. Column (3) interacts with indicators for learning style, with 'Theorist' as the omitted reference category. All regressions include baseline controls (pre-test score, gender) and school fixed effects. Robust standard errors, clustered by school, are in parentheses. *Signif. Codes:* *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The results show a remarkable consistency in the effectiveness of the tailored chatbot. Across all three models, the interaction terms between assignment to the Tailored AI (T2) and each of the heterogeneity variables are consistently small and statistically insignificant. This suggests that the tailored chatbot's ability to increase engagement is not confined to a

specific type of student; its positive effect is robust across those with different prior attitudes and learning preferences. The quality of the tool's design appears to be the primary driver of its success.

In contrast, we find some evidence that the negative engagement effect of the Generic AI (T1) is concentrated among certain student types. As shown in Column (3), the disengagement effect of the generic tool is significantly more pronounced for students with a 'Pragmatist' learning style compared to the 'Theorist' reference group (interaction coefficient of -0.111, $p < 0.10$). Furthermore, a broader analysis in Appendix E reveals that the negative effect of the Generic AI is also significantly larger for students in technical and vocational tracks. Together, these findings suggest that a poorly designed, general-purpose chatbot may exacerbate existing educational inequalities by disproportionately harming the engagement of students with certain learning styles or those in non-academic tracks.

5.3 The Impact of Chatbot Assignment on Learning (ITT Estimates)

Given that engagement is the first-order response, we now estimate the causal effect of the *offer* of each chatbot on our two primary learning outcomes. We focus on the Intent-to-Treat (ITT) effect, which is robust to the severe and differential attrition. Table 5 presents our main findings. For each outcome, we present two sets of results: our preferred conservative point estimate and non-parametric bounds on the Average Treatment Effect.

Columns 1 and 3 report the ITT point estimates from the full sample, where the outcomes for all attritors have been imputed to zero. This provides a credible lower bound on the treatment effect under the assumption that the effect of the intervention on knowledge is non-negative. Columns 2 and 4 provide a more formal robustness check by reporting the non-parametric bounds on the ATE estimated using the method of Tauchmann (2014) (See more details in Appendix F). These bounds represent the range of possible effects under worst-case assumptions about the outcomes of the attrited students, requiring only a

plausible monotonicity assumption about selection.

Table 5: The Impact of Chatbot Assignment on Learning: ITT Estimates and Lee Bounds

	Gained Financial Literacy (SD)		Knowledge Retention (SD)	
	(1) ITT (Impute 0)	(2) Lee Bounds	(3) ITT (Impute 0)	(4) Lee Bounds
Assigned to Generic AI (T1)	-0.002 (0.006)	[-0.07, 0.05]	-0.284 (0.497)	[-0.45, -0.10]
Assigned to Tailored AI (T2)	0.036*** (0.006)	[0.01, 0.11]	-0.026 (0.264)	[-0.21, 0.15]
Observations	2,440	2,440	2,440	2,440

Notes: This table reports estimates of the treatment effect on learning outcomes for the full randomized sample (N=2,440). Columns 1 and 3 report ITT point estimates from OLS regressions where the outcomes for attritors are imputed to be zero. Robust standard errors, clustered by school (58 clusters), are in parentheses. Columns 2 and 4 report non-parametric Lee (2009) bounds on the Average Treatment Effect, which account for differential attrition under a monotonicity assumption. All models include a full set of baseline controls and school fixed effects. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The results paint a clear and robust picture. Focusing first on immediate learning, our preferred point estimate in Column 1 shows that assignment to the tailored AI (T2) increased the knowledge gain score by a significant 0.036 standard deviations ($p < 0.01$). Crucially, this finding is not an artifact of our imputation choice. The Lee bounds in Column 2 for this effect are [0.01, 0.11]. Because this entire range is above zero, we can confidently conclude that the tailored chatbot had a positive causal effect on immediate learning, even accounting for the severe attrition.

For long-term knowledge retention, the effects are less clear. While the point estimates in Column 3 are negative, they are not statistically significant. The Lee bounds in Column 4 are wide and contain zero for the tailored AI. We therefore find no robust evidence that either chatbot had a lasting impact on knowledge retention at the population level. Assignment to the generic AI (T1) has no positive effect on any learning outcome; indeed, the Lee bounds for its effect on retention are strictly negative. Similar ITT analyses on psychosocial outcomes show that the tailored AI robustly boosts self-confidence, while the generic AI harms it (see Appendix F).

5.4 The Effect on the Compliers (LATE)

We now estimate the effect of the tailored chatbot on the students it successfully engaged. As established in our empirical strategy, we can only interpret a LATE for the T2 arm. Table 6 presents the 2SLS estimate for the effect of completing the T2 module on our two primary learning outcomes.

Table 6: The Effect of Treatment Completion on Learning Outcomes (LATE Estimates)

Dependent Variable:	(1) Gained Score (SD)	(2) Learning Eff. (SD)	(3) Retention (SD)
Panel A: LATE Estimates (2SLS)			
Completed Generic AI (T1)	0.1824***	0.2118***	0.2819
	(0.0392)	(0.0241)	(0.2344)
Completed Tailored AI (T2)	0.0607*	0.0618**	0.3583*
	(0.0275)	(0.0177)	(0.0901)
Panel B: Model Specification			
Observations	616	616	141
First-Stage F-statistic	> 1000	> 1000	> 1000

Notes: This table reports Two-Stage Least Squares (2SLS) estimates of the Local Average Treatment Effect (LATE). Each column is a separate regression. The endogenous variables are indicators for completing the Generic AI (T1) and Tailored AI (T2) modules, instrumented with the corresponding random assignment indicators. The first-stage relationship is mechanically strong; the F-statistics are extremely large, confirming instrument relevance. The number of observations reflects the sample for which the dependent variable is non-missing. All regressions include a full set of baseline controls and school fixed effects. Robust standard errors, clustered by school, are in parentheses. Significance codes: *** p¡0.01, ** p¡0.05, * p¡0.1.

The results show that for the compliers—students on the margin of engagement—completing the tailored chatbot module had a large and significant effect. It increased immediate knowledge gain by 0.06 standard deviations and, most importantly, increased long-term knowledge retention by 0.36 standard deviations. This shows that for the students whose behavior was changed by the intervention, the learning effects were both large and durable.

5.5 Exploring Mechanisms: Psychosocial Outcomes and Descriptive Efficiency

Finally, we explore potential mechanisms by examining outcomes measured only on the selected sample of completers. These results are descriptive and must be interpreted with extreme caution due to the severe selection bias documented in Section 3.4.

As shown in Appendix Table **??**, the most striking finding from the post-test psychosocial surveys is the powerful, opposing effect on self-confidence. Assignment to the tailored AI significantly increased self-confidence among those who completed it, while assignment to the generic AI significantly harmed it. This suggests a key channel for engagement: the tailored tool builds students' belief in their own ability, while the generic tool undermines it.

Similarly, we can descriptively examine learning efficiency for the completer sample. We find that among the selected group of students who finished, those assigned to the generic AI appear more "efficient." This is almost certainly driven by strong positive selection: only the most able and efficient students persisted with the frustrating generic tool, a finding consistent with our attrition analysis.

6 Conclusions

This paper investigates the returns to tailoring Artificial Intelligence tools for education. Using a large-scale randomized trial, we demonstrate that deep curricular integration is the critical ingredient for an effective AI-based learning tool. We show that a curriculum-tailored AI chatbot significantly boosted student engagement and completion, while a generic version actively disengaged students. This initial behavioral response is the primary mechanism driving our learning outcomes.

Our most robust causal finding is a modest but significant population-level effect on learning; the offer of the tailored chatbot increased immediate test scores by 0.036 standard deviations. The tool's full potential, however, is revealed by examining its effect on the

students whose behavior it changed. For these "compliers," completing the tailored chatbot module led to large and durable learning gains, increasing knowledge retention two months later by a significant 0.36 standard deviations. We provide evidence that a key channel for this success is the tailored tool's ability to boost student self-confidence, a stark contrast to the generic tool, which harmed it.

These findings offer a crucial insight into the educational production function in the age of AI. The distinction between the large potential effect of the tool on compliers (the LATE) and its modest population-level impact (the ITT) highlights that the primary constraint on the effectiveness of educational technology is not necessarily its pedagogical power, but its ability to solve the first-order challenge of engaging students. For policymakers and school administrators, this suggests that the promise of inexpensive, "one-size-fits-all" AI solutions may be illusory. Our results indicate that deep curricular integration is essential to unlock the technology's potential and, crucially, to avoid the unintended consequence of disengaging the very students the technology is meant to help.

Our study has limitations. Our definition of a "tailored" chatbot combines both content-specificity and pedagogical adaptivity, and our design does not allow us to disentangle these two components. Furthermore, our results are from a single, albeit complex, subject area (taxation) in one country, and future work should explore whether these returns to tailoring hold across different domains and student populations. Nonetheless, our findings underscore a fundamental principle: for educational technology to be effective, it must first be used. Future research should move beyond asking if technology works, and instead focus on quantifying the returns to specific design features like contextualization, as this appears to be the central margin for generating productivity gains in education.

References

Angrist, J. D., & Pischke, J.-S. (2009a). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2009b). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Aucejo, E. M., & Romano, T. F. (2016). Assessing the effect of school days and absences on test score performance. *Economics of Education Review*, *55*, 70–87. doi: 10.1016/j.econedurev.2016.08.007

Bai, Y., Mo, D., Zhang, L., Boswell, M., & Rozelle, S. (2016). The impact of integrating ict with teaching: Evidence from a randomized controlled trial in rural schools in china. *Computers And Education*, *96*, 1–14. doi: 10.1016/j.compedu.2016.02.005

Cacault, M. P., Hildebrand, C., Laurent-Lucchetti, J., & Pellizzari, M. (2021). Distance learning in higher education: Evidence from a randomized experiment. *Journal of the European Economic Association*, *19*(4), 2322–2372. doi: 10.1093/jeea/jvaa060

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research*, *66*(3), 227–268. doi: 10.3102/00346543066003227

Cunningham, S. (2018). *Causal Inference: The Mixtape*. (Online manuscript version. Later published by Yale University Press in 2021.)

De Witte, K., De Beckker, K., & Holz, O. (2020, June 18). Financial education in flanders (belgium). In K. De Witte, O. Holz, & K. De Beckker (Eds.), *Financial education* (pp. 67–85). Germany: Waxmann Verlag GMBH.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics. In *Handbook of development economics* (Vol. 4, pp. 3895–3962). Elsevier.

Figlio, D., Rush, M., & Yin, L. (2013). Is it live or is it internet? experimental estimates of the effects of online instruction on student learning. *Journal of Labor Economics*, *31*(4), 763–784. doi: 10.1086/669930

Hanushek, E. (2003). The failure of input-based schooling policies. *Economic Journal*, *113*(485), F64-F98. Retrieved from https://EconPapers.repec.org/RePEc:ecj:econjl:v:113:y:2003:i:485:p:f64-f98

Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., & Lee, A. (2024). Effective and Scalable Math Support: Evidence on the Impact of an AI- Tutor on Math Achievement in Ghana. *arXiv.org*. Retrieved from https://arxiv.org/abs/2402.09809 doi: 10.48550/ARXIV.2402.09809

Honey, P., & Mumford, A. (1986). *The manual of learning styles*. Peter Honey. Retrieved from https://books.google.be/books?id=4TV-twAACAAJ

Jaume, D., & Willén, A. (2019). The long-run effects of teacher strikes: Evidence from argentina. *Journal of Labor Economics*, *37*(4), 1097–1139. doi: 10.1086/703138

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2024). Ai tutoring outperforms active learning. *Research Square*. (Preprint) doi: 10.21203/rs.3.rs-3965934/v1

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*(1), 83–119.

Lusardi, A., & Mitchell, O. S. (2014, March). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature*, *52*(1), 5–44. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jel.52.1.5 doi: 10.1257/jel.52.1.5

OCDE. (2024). *Resultados pisa 2022 (volumen iv): ¿qué tan inteligentes financieramente son los estudiantes?* París: Publicaciones de la OCDE. Retrieved from https://doi.org/10.1787/5a849c2a-es doi: 10.1787/5a849c2a-es

OECD. (2020). *OECD/INFE 2020 International Survey of Adult Financial Literacy*. Retrieved 2023-10-27, from https://www.oecd.org/content/dam/oecd/en/publications/reports/2020/06/oecd-infe-2020-international-survey-of-adult-financial-literacy_bbad9b27/145f5607-en.pdf (Accessed: 2023-10-27)

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). A manual for the

use of the motivated strategies for learning questionnaire (mslq) [Computer software manual]. Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning, University of Michigan.

Pressley, T. (2021). Factors contributing to teacher burnout during covid-19. *Educational Researcher*, *50*(5), 325–327. doi: 10.3102/0013189X211004683

Sales, A. C., & Pane, J. F. (2020). *Student log-data from a randomized evaluation of educational technology: A causal case study* (Tech. Rep.). RAND Corporation. (Published in Journal of Research on Educational Effectiveness, 13:2, 237-259) doi: 10.1080/19345747.2019.1678257

Schaufeli, W. B., Salanova, M., González-Romá, V., & Bakker, A. B. (2002). The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. *Journal of Happiness Studies*, *3*(1), 71–92. doi: 10.1023/A:1015630930326

Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. In J. Weinman, S. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. causal and control beliefs* (pp. 35–37). Windsor, UK: NFER-NELSON.

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2019). Understanding teacher shortages: An analysis of teacher supply and demand in the united states. *Education Policy Analysis Archives*, *27*(35). doi: 10.14507/epaa.27.3626

Tauchmann, H. (2014). Lee (2009) treatment-effect bounds for nonrandom sample selection. *The Stata Journal*, *14*(4), 884-894. Retrieved from `https://doi.org/10.1177/1536867X1401400411` doi: 10.1177/1536867X1401400411

Tlili, A., Saqer, K., Salha, S., & Huang, R. (2025, jan 17). Investigating the effect of artificial intelligence in education (AIEd) on learning achievement: A meta-analysis and research synthesis. *Information Development*. Retrieved from `http://dx.doi.org/10.1177/02666669241304407` doi: 10.1177/02666669241304407

Vlaamse Regering. (n.d.). *Voorontwerp van decreet tot wijziging van de codex secundair onderwijs van 17 december 2010, wat betreft de modernisering van de structuur en*

de organisatie van het secundair onderwijs (Voorontwerp van decreet No. VR 2018 0202 DOC.0094/3BIS). Retrieved 2024-05-21, from https://www.klasse.be/73458/nieuw-model-studieaanbod-secundair/

Wang, X., Huang, R. T., Sommer, M., Pei, B., Shidfar, P., Rehman, M. S., ... Martin, F. (2024, may 15). The Efficacy of Artificial Intelligence-Enabled Adaptive Learning Systems From 2010 to 2022 on Learner Outcomes: A Meta-Analysis. *Journal of Educational Computing Research*, *62*(6), 1348–1383. Retrieved from http://dx.doi.org/10.1177/07356331241240459 doi: 10.1177/07356331241240459

Wu, R., & Yu, Z. (2023, may 3). Do AI chatbots improve students learning outcomes? Evidence from a metaanalysis. *British Journal of Educational Technology*, *55*(1), 10–33. Retrieved from http://dx.doi.org/10.1111/bjet.13334 doi: 10.1111/bjet.13334

A Appendix Data: Baseline Balance Check

Table 7: Baseline Balance Check: Categorical Variables

Variable	(1) Control (%)	(2) Generic AI (%)	(3) Tailored AI (%)	(4) p-value (χ^2 test)
Gender				0.695 ($\chi^2(4) = 2.22$)
Female	50.45	48.41	47.00	
Male	49.55	51.59	52.99	
School Type				0.635 ($\chi^2(8) = 6.20$)
General Secondary (ASO)	68.03	65.82	64.27	
Technical Secondary (TSO)	29.30	31.65	32.00	
Vocational Secondary (BSO)	2.04	1.93	2.44	
Secondary Education in the Arts (KSO)	0.38	0.36	0.51	
Other	0.255	0.24	0.77	
Secondary School Field of Study				0.593 ($\chi^2(12) = 10.26$)
Arts & Sports	1.665	2.182	0.907	
Care & Social Studies	10.371	8.848	9.974	
Economics & Business	20.871	20.970	20.337	
Humanities & Languages	22.663	20.606	20.596	
Vocational & Applied Skills	1.152	1.091	1.425	
STEM	36.748	38.061	40.285	
Others	6.530	8.242	6.477	
Last Dutch Grade				0.290 ($\chi^2(8) = 9.66$)
Under 50%	2.80	2.41	2.70	
50% to 59%	13.76	14.44	12.08	
60% to 69%	40.26	42.48	40.62	
70% to 79%	34.90	29.84	35.09	
Over 80%	8.28	10.83	9.51	
Last Math Grade				0.440 ($\chi^2(8) = 7.94$)
Under 50%	7.52	6.86	8.61	
50% to 59%	23.06	22.86	20.69	
60% to 69%	30.06	32.49	32.01	
70% to 79%	23.95	23.47	26.35	
Over 80%	15.41	14.32	12.34	
Language at home				0.634 ($\chi^2(4) = 2.56$)
Dutch	83.43	83.27	81.10	
French	5.61	6.38	6.43	
Other	10.95	10.34	12.47	
Highest Parents' Educational Level				0.554 ($\chi^2(6) = 4.92$)
Higher education degree	68.92	68.95	68.89	
Secondary education	18.22	16.73	15.68	
No secondary education	3.44	3.01	4.24	
Unknown	9.43	11.31	11.18	
Frequency of Asking Teachers for Help				0.643 ($\chi^2(6) = 6.17$)
Always	0.30	0.48	0.39	
Never	11.46	12.64	13.24	
Often	8.79	8.42	7.58	
Rarely	38.60	38.15	36.50	
Sometimes	41.15	40.31	42.29	
Learning Style (Honey & Mumford, 1986)				0.094 ($\chi^2(6) = 10.82$)
Activist	28.79	33.09	32.13	
Pragmatist	25.99	23.35	24.04	
Reflector	19.36	18.41	22.37	
Theorist	25.86	25.15	21.47	
Observations	799	870	771	

Notes: This table reports the fraction of students in various categorical groups at baseline for the full randomized sample (N=2,394). Column 4 reports the p-value from a Pearson's χ^2 test for the independence of the variable and treatment assignment status across all three groups. No test yields a p-value significant at the 10% level, confirming successful randomization across observable categorical characteristics. For brevity, some categorical variables from the original table have been omitted but show similar balance.

B Appendix: Experimental Materials

Pre-Test Questionnaire

- A1. What is your full first and last name? (For example: Johnson John)
- A2. You are a ...
  - Boy
  - Girl
  - X
- A3. Municipality/City of your school:
- A4. Name of your school:
- A5. Which study program do you follow? (e.g. Economics-Mathematics, Economics-Modern Languages or Business Studies)
- A6. In which type of education are you in school?
  - General Secondary Education (ASO)
  - Technical Secondary Education (TSO)
  - Vocational Secondary Education (BSO)
  - Art Secondary Education (KSO)
  - Other
- B1. What was your last grade for Dutch at the end of last school year?
  - Less than 50%
  - 50% or more, but less than 60%
  - 60% or more, but less than 70%

  - 70% or more, but less than 80%
  - More than 80%

- B2. What was your last grade for mathematics at the end of last school year?
  - Less than 50%
  - 50% or more, but less than 60%
  - 60% or more, but less than 70%
  - 70% or more, but less than 80%
  - More than 80%

- B3. Which language do you speak most at home?
  - Dutch
  - French
  - Other

- B4. What is the highest educational level of your parents (mother or father) who live at your home?
  - No secondary education/secondary school not completed
  - Diploma of secondary education/secondary school
  - College/university degree or higher
  - I don't know

- B5. How many (step)brothers and (step)sisters still live at home?
  - 0
  - 1
  - 2

  - 3 or more

- B6. How motivated are you to perform well at school?
  - Very unmotivated
  - Unmotivated
  - Neutral
  - Motivated
  - Very motivated

- B7. How often do you ask your teachers for help with schoolwork or studying?
  - Never
  - Rarely
  - Sometimes
  - Often
  - Always

- C1. How much time do you spend on social media and the internet (such as watching videos, surfing, or chatting) on a typical day?
  - 0-1 hour
  - 1-2 hours
  - 2-4 hours
  - More than 4 hours

- C2. How do you feel about the following activities?
  - I use AI assistants a lot (such as ChatGPT or Gemini)

  - I already know a lot about financial concepts such as taxes, budgeting, saving and investing
  - I find lessons about financial concepts interesting.
  - AI tools can help me study.

- D1. Answer the following question based on your preferences:
  - I like to learn by doing experiments and trying things out myself.
  - I am not afraid to take risks and try new things when I learn.
  - I like to think carefully about things before I do them.
  - I learn best when I have time to think about my experiences.
  - I want to understand how things work and why things are the way they are.
  - I like to analyze information and put the pieces together to figure things out.
  - I want to learn things that I can actually use in real life.
  - I like clear instructions and know exactly what to do.

- E1. Jan pays €1000 tax on an income of €5000. What is his average tax rate (tax percentage)?
  - 10%
  - 20%
  - 25%
  - 50%
  - I don't know

- E2. Answer the following question based on your preferences:
  - Do you think taxes are fair in your country?

  - Do you think people in your country know much about taxes?
  - Taxes are essential for funding public services.
  - In general, I feel comfortable performing calculations with numbers
  - I expect AI to help me learn about taxes.

- E3. Peter has an income of €2200 per month. Bart earns €3800 per month. Calculate the pay gap
  - 173
  - 58
  - 43
  - 73
  - I don't know

- E4. Which tax system leads to the most equal income distribution?
  - Degressive tax system
  - Proportional tax system (flat tax)
  - Progressive tax system
  - None of the above
  - I don't know

- E5. A freelancer earns €57,000 gross per year. In a tiered progressive tax system with the following brackets, what is the tax payable (round to whole euros)?
  - Bracket Income bracket (gross per year) Tax rate (%)
  - 1 €0-€20,000 25
  - 2 €20,000-€40,000 40

  - 3 over €40,000 53
  - €25010
  - €26790
  - €22010
  - €30210
  - I don't know

- E6. Ann earns €43,000 gross per year. How much would she have left if the tax rate is 30%?
  - 30100
  - 26667
  - 14333
  - 12900
  - I don't know

- E7. Which factor has the LEAST direct influence on the calculation of income tax?
  - Professional costs
  - Number of dependent children
  - The national average wage
  - Tax-free amount
  - I don't know

- E8. Mattice has a gross annual income of €43,000 and pays €26794 in taxes. What is the average tax rate?
  - 62%

  - 23%
  - 160%
  - 165%
  - I don't know

  "'latex

- E9. A self-employed person with an income of €50,000 is considering taking on an extra assignment worth €10,000. Which of the following statements is most correct regarding the impact of this additional income on her tax burden?
  - In a globally progressive system, the extra assignment would always result in a higher net income.
  - In a tiered progressive system, the tax rate on the additional income would be identical to that on the initial income.
  - In a degressive system, the total average tax rate on the income would fall after the extra assignment.
  - I don't know.

- F1. Answer the following question based on your preferences:
  - In general, I enjoy learning new subjects, even if they are not directly among my interests.
  - I think knowledge about financial matters can be useful in the future.
  - I am usually open to extra teaching material or tools to help me learn.
  - I like to discover new ways to learn.
  - I expect AI can help me learn about taxes.
  - I like working with computers.

  - I like working with AI tools.

- F2. Answer the following question based on your preferences:
  - I like to try out new digital tools if they can be useful for my studies.
  - I usually don't find it difficult to work with new (online) tools.
  - If I have to use a new digital tool, I am usually willing to put in some extra time to learn it.

- F3. Answer the following question based on your preferences:
  - I often make a plan or schedule before I start my schoolwork.
  - While learning, I pay attention to whether I really understand the material and adjust my approach if not.
  - If I don't immediately understand something, I try to find out what I can do better or differently.

- F4. Answer the following question based on your preferences:
  - I usually find it important to fully commit to my schoolwork.
  - I can usually concentrate well when I am working on an assignment.
  - I often feel like finding out more about the topics covered in class.

- F5. Answer the following question based on your preferences:
  - If something is complicated, I believe I can understand it if I try my best.
  - In general, I feel confident when I start a new challenge for school.

- F6. Answer the following question based on your preferences:
  - I sometimes feel nervous if I don't know what to expect from a subject or lesson topic.

  - I look forward to the challenge of learning something new, even though it may be difficult.

Post-Test Questionnaire

- A1. What is your full name? (Example: Jansen Jan)
- A2. To which group were you assigned?
  - Learning path group 1
  - Learning path group 2
  - Learning path group 3
- A3. You are a ...
  - Boy
  - Girl
  - X
- A4. Municipality/City of your school:
- A5. Name of your school:
- A6. Which study program do you follow? (e.g. Economics-Mathematics, Economics-Modern Languages or Business Studies)
- A7. In which form of education do you follow lessons?
  - General Secondary Education (ASO)
  - Technical Secondary Education (TSO)
  - Vocational Secondary Education (BSO)
  - Art Secondary Education (KSO)

  - Other

- A8. Where did you follow the digital lesson?
  - In the regular class with my economics teacher
  - In the regular class but not with my usual teacher
  - In study
  - At home

- B1. Lisa pays €1800 in taxes on an income of €7200. What is her average assessment rate (in percentage)?
  - 30%
  - 25%
  - 20%
  - 35%
  - I don't know

- B2. A junior employee earns €2100 per month. A senior manager earns €5100 per month. Calculate the wage gap (rounded to the nearest whole number).
  - 243
  - 58
  - 143
  - 41
  - I don't know

- B3. A self-employed person earns €38,000 gross per year. With a tiered progressive tax system with the following brackets, what is the tax payable (round to the nearest whole number):

  - Bracket Income bracket (gross per year) Tax rate (%)
  - 1 0-€22,000 25
  - 2 €22,000-€42,000 40
  - 3 above €42,000 50
  - €22800
  - €6180
  - €11900
  - €15200
  - I don't know

- B4. Which tax system leads to the most equal income distribution?
  - Degressive tax system
  - Proportional tax system (flat tax)
  - Progressive tax system
  - None of the above
  - I don't know

- B5. Thomas has a gross annual salary of €48,000. If the tax rate is 32%, how much does he have left after taxes?
  - €32640
  - €15360
  - €40000
  - €53500
  - I don't know

- B6. Which factor has the LEAST direct influence on an individual's income tax?
  - The professional costs
  - The average national wage
  - Number of dependent children
  - The tax-free sum
  - I don't know
- B7. Paul has a gross annual income of €52,000 and pays €19,240 in taxes. What is his average assessment rate?
  - 37%
  - 63%
  - 270%
  - 165%
  - I don't know
- B8. A self-employed person with an income of €50,000 is considering taking on an extra assignment worth €10,000. Which of the following statements is most correct regarding the impact of this additional income on her tax burden?
  - In a globally progressive system, the extra assignment would always result in a higher net income.
  - In a tiered progressive system, the tax rate on the additional income would be identical to that on the initial income.
  - In a degressive system, the total average tax rate on the income would fall after the extra assignment.
  - I don't know.

- B9. Answer the following question based on your preferences:
  - Do you think taxes are fair in Belgium?
  - Do you think people know a lot about taxes?
  - Taxes are essential for the financing of public services.
  - In general, I have no problem performing calculations with numbers
  - I expect AI can help me learn about taxes.
- C1. Answer the following question based on your preferences:
  - I like to learn new knowledge about taxes.
  - If I understand more about taxes, this will be useful to me in the future.
  - I find it interesting to learn how tax rules work.
  - The digital lesson I took made me motivated about the subject of taxes.
- C2. Answer the following question based on your preferences:
  - The digital lesson I took helps me to better understand the subject matter of taxes.
  - I feel that the digital lesson I took supported my learning process about taxes.
  - The digital lesson I took is simple and intuitive to use.
  - I would like to use such a digital lesson more often in the future for other subjects.
  - I found that there was a lot of repetition in the digital lesson.
  - I was able to follow the instruction well in the digital lesson.
- C3. Answer the following question based on your preferences:
  - I often make a plan or schedule before I start my schoolwork.

  - While learning, I pay attention to whether I really understand the material and adjust my approach if not.
  - If I don't immediately understand something, I try to find out what I can do better or differently.

- C4. Answer the following question based on your preferences:
  - I feel enthusiastic when I can learn more about the theme of taxes.
  - I am completely absorbed in the activities related to tax subjects.
  - I have a lot of energy when I have to learn about taxes.
  - I was able to learn a lot in the digital lesson.
  - I was able to concentrate well in the digital lesson.

- C5. Answer the following question based on your preferences:
  - When I made a mistake in reasoning, I was able to figure out how this came about in the digital lesson.
  - If I notice that I do not understand something about taxes, I could find an answer to my questions in the digital lesson.
  - Thanks to the digital lesson, I think about how I can apply what I have learned about taxes in daily life.

- C6. Answer the following question based on your preferences:
  - I feel enthusiastic when I can learn more about the theme of taxes.
  - I am completely absorbed in the activities related to tax subjects.
  - I have a lot of energy when I have to learn about taxes.
  - I was able to learn a lot in the digital lesson.
  - I was able to concentrate well in the digital lesson.

C Appendix: Additional Tables

C.1 Statistical Tests for Attrition Analysis

Table 8 provides the statistical foundation for the attrition analysis presented in Section 3.4. It reports the p-values from two-sample t-tests comparing the means of baseline characteristics for students who completed the post-test versus those who did not, within each of the three experimental arms.

Table 8: Statistical Tests for Differences in Baseline Characteristics Between Completers and Non-Completers

	P-value of Difference (Completer vs. Non-Completer)		
Baseline Variable	Control (T0)	Generic AI (T1)	Tailored AI (T2)
Pre-Test Score	0.008***	<0.001***	0.082*
Attitude & Motivation	0.164	0.111	0.532
AI Attitude & Motivation	0.048**	0.653	0.032**
Learning Experience	<0.001***	0.077*	0.717
Self-Regulation	0.713	0.848	0.646
Engagement & Commitment	0.387	0.527	0.533
Self-Confidence	0.283	0.007***	0.913
Emotional Factors	0.473	0.284	0.758

Notes: This table reports the p-values from two-sample t-tests comparing the means of baseline characteristics for students who completed the post-test versus those who did not, within each treatment arm. The table supports the analysis in Section 3.4.
*Significance codes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.*

D Appendix: Diagnostic Analysis, The Impact of Selection Bias on Naive Estimates

In this section, we provide a diagnostic analysis to illustrate the severe bias that arises from failing to account for non-random attrition. While our main analysis relies on ITT estimates with imputation and Lee Bounds on the full sample, examining the naive estimates on the subsample of completers is instructive. It reveals the potential of the interventions for the selected group of students who use them and highlights the critical importance of our primary empirical strategy.

Table 9 presents two sets of ITT estimates. Panel A reports our conservative lower-

bound estimates on the full sample with zero-imputation for our validly imputable learning outcomes. Panel B reports naive OLS estimates on the unrepresentative subsample of students who completed the post-test or follow-up test.

Table 9: The Impact of AI Assignment on Learning Outcomes: ITT Estimates

	Full Sample (Imputed)			**Completer Sample**		
Dependent Variable:	(1) Gain Score (SD)	(2) Learning Eff. (SD)	(3) Retention (SD)	(4) Gain Score (SD)	(5) Learning Eff. (SD)	(6) Retention (SD)
Panel A: Treatment Effects						
Assigned to Generic AI (T1)	-0.0024	0.0333***	-0.2835	0.0625***	0.2584***	0.2685
	(0.0064)	(0.0102)	(0.4970)	(0.0178)	(0.0555)	(0.2214)
Assigned to Tailored AI (T2)	0.0356***	0.0260	-0.0256	0.0127***	0.0465**	0.3872***
	(0.0062)	(0.0251)	(0.2640)	(0.0048)	(0.0182)	(0.0643)
Panel B: Model Specification						
Observations	2,440	2,440	2,440	616	616	141
R-squared	0.035	0.258	0.325	0.300	0.235	0.153
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Baseline Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports Intent-to-Treat (ITT) estimates from OLS regressions. All dependent variables are standardized to have a mean of 0 and a standard deviation of 1 relative to the control group. Columns (1)-(3) use the full randomized sample (N=2,440). Outcomes for students who attrited are imputed to be zero. These are our primary, conservative lower-bound estimates. Columns (4)-(6) use the subsample of students who completed the relevant survey (post-test for Gain Score and Learning Efficiency, N=616; follow-up test for Retention, N=141). These estimates are conditional on completion and are presented to illustrate the impact of selection. All regressions include a full set of baseline controls (pre-test score, gender, parental education, prior grades) and school fixed effects. Robust standard errors, clustered by school, are in parentheses. Significance codes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The contrast between Panel A and Panel B is stark and demonstrates the importance of our identification strategy. The most dramatic difference appears for knowledge retention. In our primary analysis (Panel A, Column 2), we find no significant population-level effect of the tailored AI on retention. However, the naive estimate on the selected completer sample (Panel B, Column 4) shows a massive and highly significant effect of 0.387 standard deviations.

This large discrepancy reveals two things. First, it highlights the **potential** of the tailored chatbot: for the motivated and persistent students who complete the module, the tool is exceptionally effective at fostering durable, long-term learning. This is a crucial finding for understanding the pedagogical power of the intervention. Second, it underscores the severity of the selection bias. The large effect in Panel B is realized only by a non-random

subset of students. Our more conservative and credible ITT and LATE analyses in the main text correctly account for this selection to estimate the true causal effects for the broader population and for the marginal student, respectively. The divergence between these panels provides a powerful, non-parametric illustration of why such methods are essential.

E Appendix: Heterogeneity Analysis

This section provides the full set of results for the heterogeneity analysis summarized in Section 5.2. Table 10 presents coefficients from eight separate estimation where treatment assignment is interacted with a different baseline student characteristic.

Table 10: Heterogeneous Effects of Chatbot Assignment on Module Completion (Full Analysis)

Variable	(1) By Gender	(2) By TSO/ BSO	(3) By GSO Track	(4) By Baseline Score	(5) By Teacher Shortage	(6) By Help-Seeking	(7) By Parent Education	(8) By Home Language
Treatment Main Effects								
Assigned to T1	-0.041*	-0.006	-0.090**	-0.040	-0.029	-0.028	-0.073**	-0.034
	(0.023)	(0.021)	(0.037)	(0.029)	(0.021)	(0.021)	(0.037)	(0.023)
Assigned to T2	0.130***	0.133***	0.138***	0.156***	0.134***	0.127***	0.161***	0.120***
	(0.033)	(0.032)	(0.044)	(0.040)	(0.028)	(0.026)	(0.045)	(0.028)
Interaction Effects								
T1 × Subgroup	0.016	-0.081**	0.087**		-0.053	-0.049	0.058	0.024
	(0.029)	(0.040)	(0.040)		(0.055)	(0.059)	(0.047)	(0.050)
T2 × Subgroup	-0.000	-0.008	-0.010		-0.056	0.047	-0.046	0.070
	(0.043)	(0.053)	(0.053)		(0.051)	(0.092)	(0.053)	(0.067)
T1 × Low Score				0.008				
				(0.043)				
T2 × Low Score				-0.039				
				(0.054)				
T1 × High Score				0.030				
				(0.056)				
T2 × High Score				-0.095				
				(0.074)				
Model Information								
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,440	2,440	2,440	2,440	2,440	2,440	2,440	2,434

*Notes: This table reports coefficients from eight separate estimations. The dependent variable is an indicator for module completion. Each column interacts the treatment assignment variables with a different baseline characteristic. "Subgroup" refers to the characteristic in the column header (e.g., Female in Col 1, TSO/BSO in Col 2). All models include baseline controls (score, gender) and school fixed effects. Robust standard errors, clustered by school, are in parentheses. Significance codes: *** p¡0.01, ** p¡0.05, * p¡0.1.*

F Appendix: Robustness to Alternative Imputation Assumptions

Our primary analysis in the main text uses a conservative zero-imputation for attritors to establish a lower bound on the Intent-to-Treat (ITT) effect. While standard, the resulting estimates are sensitive to this specific assumption. To assess the robustness of our conclusions, we perform a Manski-style bounds analysis by re-estimating the ITT effect under different, plausible assumptions about the performance of students who attrited. This appendix presents the results for our primary learning outcome (immediate knowledge gain) and all measured psychosocial outcomes.

F.1 A.1 Sensitivity Analysis for Immediate Knowledge Gain

Table 11 presents the sensitivity analysis for our main learning outcome. The results confirm the robustness of our core finding. Our main lower-bound specification is in Column (1) for reference, where the estimated effect of the tailored AI (T2) on knowledge gain is a significant 0.035 standard deviations. Column (2), which assumes attritors would have achieved the minimum score observed among any completer, yields an even larger positive effect of 0.076 standard deviations. In Column (3), where we assume attritors would have performed at the median level, the effect of the tailored AI remains positive and statistically significant (0.008 SDs). The effect only turns negative under the extreme and unrealistic assumption in Column (4) that all 1,673 attritors would have been top-performers. This analysis strongly reinforces our conclusion that the tailored AI produced a genuine, positive causal effect on learning for the student population.

Table 11: Sensitivity of ITT Estimates to Imputation Assumptions: Immediate Knowledge Gain

Dependent Variable:	Immediate Knowledge Gain (Standardized)			
	(1)	(2)	(3)	(4)
	Impute 0	Impute Min	Impute Median	Impute Max
Assigned to Generic AI (T1)	-0.0018	-0.0222**	0.0117	0.0456***
	(0.0060)	(0.0055)	(0.0063)	(0.0073)
Assigned to Tailored AI (T2)	0.0349***	0.0755***	0.0078**	0.0598***
	(0.0058)	(0.0118)	(0.0020)	(0.0082)
Observations	2,440	2,440	2,440	2,440
R-squared	0.013	0.026	0.003	0.031
School Fixed Effects	Yes	Yes	Yes	Yes

Notes: This table examines the sensitivity of our main ITT estimates to different assumptions about the outcomes of the 1,673 students who attrited. Each column reports results from an OLS regression on the full randomized sample (N=2,440) with a different imputed value for attritors. Column (1) imputes a value of zero. Columns (2)-(4) impute the minimum, median, and maximum value of the outcome variable, respectively, calculated from the subsample of completers. All regressions include school fixed effects. Robust standard errors, clustered by school, are in parentheses. Significance codes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

F.2 A.2 Sensitivity Analysis for Psychosocial Outcomes

Table 12 reports the same sensitivity analysis for all six of our measured psychosocial outcomes. In stark contrast to the results for knowledge gain, the estimated effects on these non-cognitive dimensions are highly sensitive to the imputation assumption, underscoring the challenge of drawing firm conclusions about the program's broader psychological impacts at the population level.

A consistent pattern emerges across most panels. For the tailored AI (T2), the estimated effects are often statistically insignificant or even negative under our conservative zero- and median-imputation scenarios (Columns 1 and 3). However, for many outcomes—including Engagement, Attitude, and Self-Confidence—the effect becomes large, positive, and statistically significant when we impute the minimum observed value (Column 2). This pattern suggests that while the tailored AI may have had positive psychosocial benefits for some students, particularly those who might have otherwise reported very negative outcomes, these effects do not robustly extend to the entire student population.

The most robust finding in this domain relates to Self-Confidence & Self-Efficacy (Panel

C). Here, the tailored AI shows a positive and at least marginally significant effect across the first three, most plausible imputation scenarios. This provides suggestive evidence that offering the tailored AI tool may have had a genuine positive impact on student self-confidence across the full sample, a finding that contrasts sharply with the consistently negative estimates for the generic AI. Overall, this analysis highlights the value of our conservative lower-bound approach and suggests that the program's most robust and widespread impact was on cognitive learning rather than on non-cognitive attitudes.

Table 12: Sensitivity of ITT Estimates for Psychosocial Outcomes

	(1) Impute 0	(2) Impute Min	(3) Impute Median	(4) Impute Max
Panel A: Gain in Engagement & Commitment				
Assigned to Generic AI (T1)	0.0078**	-0.1623***	0.0011	0.1745***
	(0.0024)	(0.0156)	(0.0029)	(0.0113)
Assigned to Tailored AI (T2)	-0.0374	0.2702**	-0.0253	-0.3389***
	(0.0474)	(0.0771)	(0.0485)	(0.0254)
Panel B: Gain in Attitude & Motivation				
Assigned to Generic AI (T1)	-0.0101	-0.1602***	-0.0226	0.1400***
	(0.0101)	(0.0217)	(0.0110)	(0.0040)
Assigned to Tailored AI (T2)	-0.0377***	0.2360***	-0.0149	-0.3114***
	(0.0080)	(0.0385)	(0.0105)	(0.0232)
Panel C: Gain in Self-Confidence & Self-Efficacy				
Assigned to Generic AI (T1)	-0.0168	-0.1641***	-0.0249*	0.1797***
	(0.0111)	(0.0198)	(0.0115)	(0.0049)
Assigned to Tailored AI (T2)	0.0466*	0.3230***	0.0619**	-0.3221***
	(0.0174)	(0.0508)	(0.0193)	(0.0275)
Panel D: Gain in Learning Experience & User Experience				
Assigned to Generic AI (T1)	0.0011	-0.1657***	0.0011	0.1263***
	(0.0068)	(0.0197)	(0.0068)	(0.0046)
Assigned to Tailored AI (T2)	0.0327	0.3368***	0.0327	-0.1954**
	(0.0247)	(0.0213)	(0.0247)	(0.0476)
Panel E: Gain in Self-Regulation & Metacognition				
Assigned to Generic AI (T1)	0.0007	-0.1630***	0.0007	0.1644***
	(0.0067)	(0.0164)	(0.0067)	(0.0055)
Assigned to Tailored AI (T2)	-0.0279**	0.2792***	-0.0279**	-0.3351***
	(0.0074)	(0.0302)	(0.0074)	(0.0443)
Panel F: Gain in Emotional & Psychological Factors				
Assigned to Generic AI (T1)	0.0171***	-0.1580***	-0.0129***	0.1723***
	(0.0024)	(0.0122)	(0.0017)	(0.0143)
Assigned to Tailored AI (T2)	-0.0689	0.2504**	-0.0142	-0.3518***
	(0.0336)	(0.0662)	(0.0388)	(0.0200)
Observations	2,407	2,407	2,407	2,407

Notes: This table reports the sensitivity analysis for changes in all measured psychosocial outcomes. Each panel represents a different dependent variable. See notes for Table 11 for details on the column specifications and methodology. All regressions include school fixed effects. Robust standard errors, clustered by school, are in parentheses. Significance codes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.